eric

Enteropathogen Resource Integration Center

Bioinformatics Resource Center

John Greene, Ph.D., SRA International

&

Nicole Perna, Ph.D., U. Wisconsin-Madison

BRC Kickoff Meeting
October 13, 2004

Enteropathogen Resource Integration Center

- ERIC will focus on integration of data from 5 enteropathogens:
  - Diarrheagenic *E. coli*
  - *Shigella* spp.
  - *Salmonella* spp.
  - *Yersinia enterocolitica*
  - *Yersinia pestis*
- Partnership between personnel at the Genome Center of Wisconsin and SRA International, Rockville MD

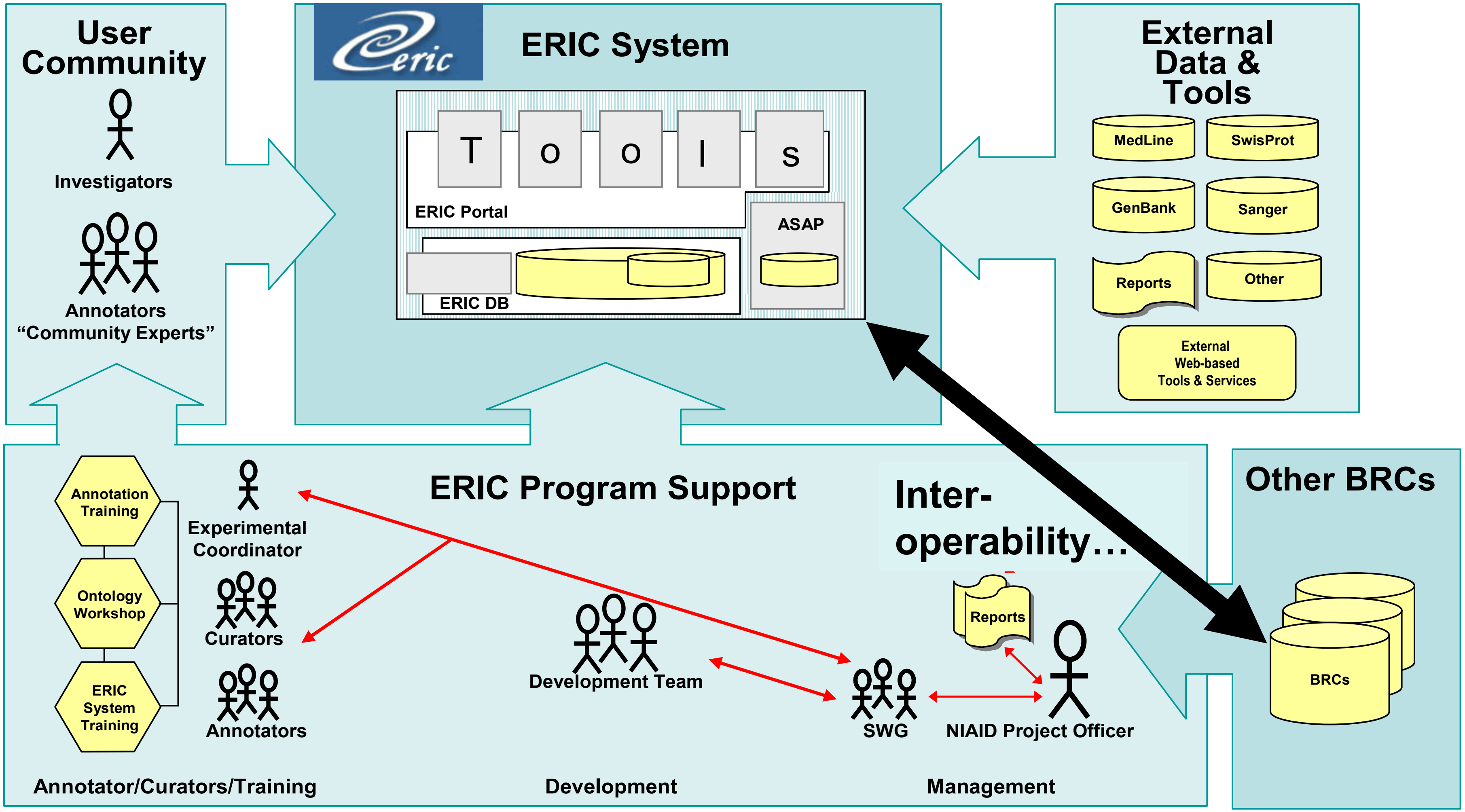
http://www.ericbrc.org

- ERIC is in early development – please continue to check back as new features are added

- Pathogen-centric, not technology-centric vision

- Community annotation via ASAP is functional NOW

- Suggestions, comments, criticism are welcome by e-mail at info@ericbrc.org

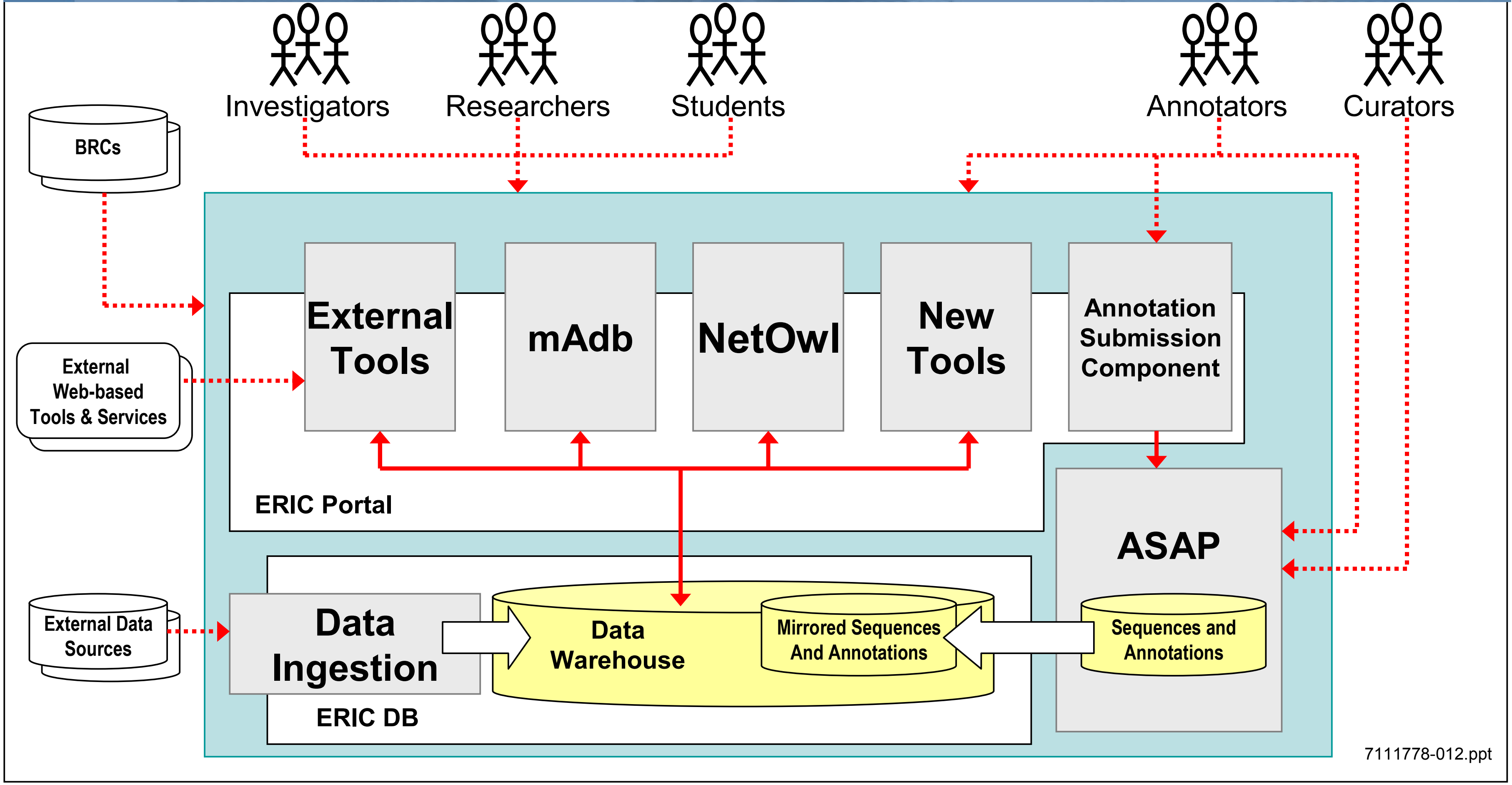


7111778-026.ppt

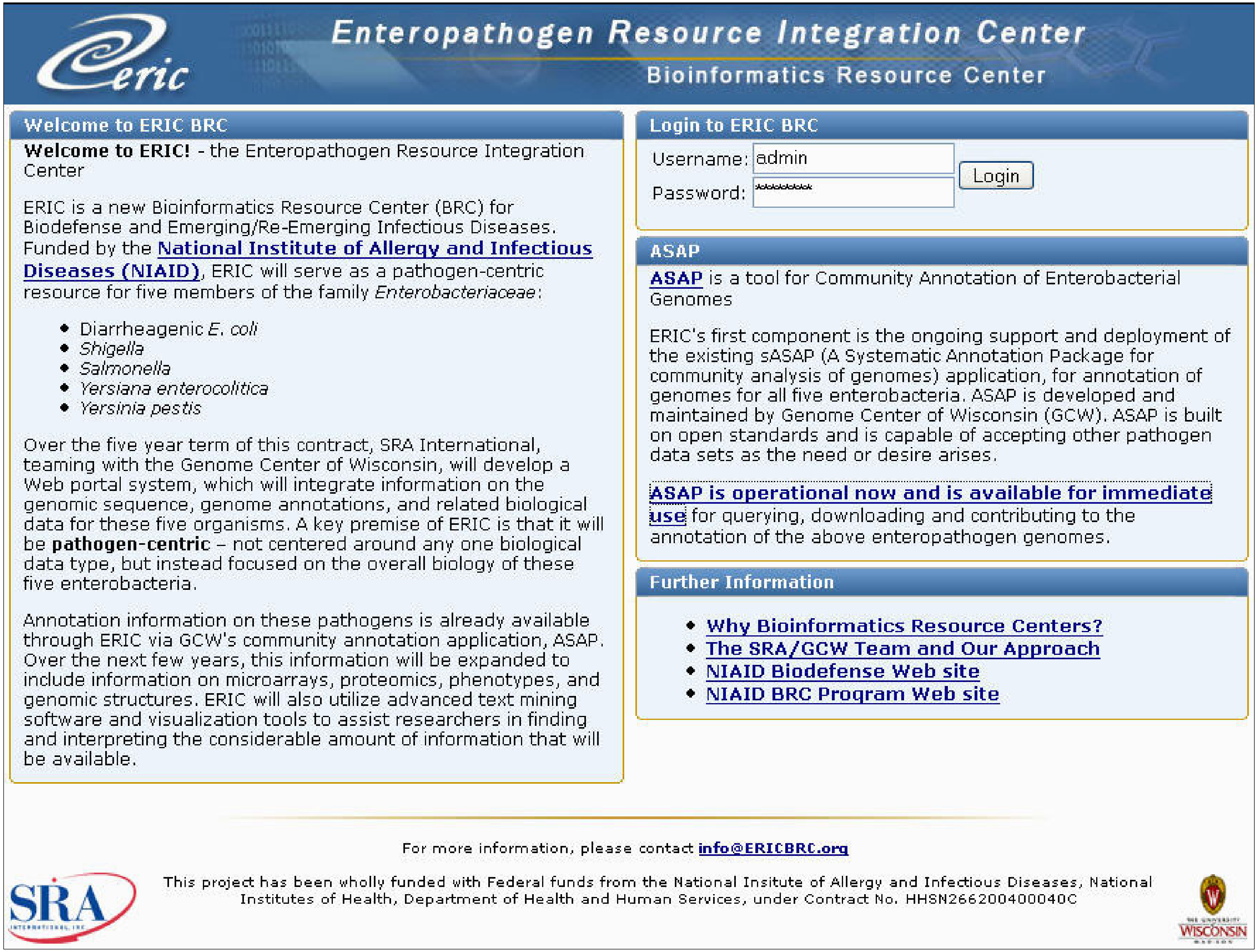
ERIC will have many inputs, but the **most important inputs will come from the user community -** suggestions for tools, participation in annotation and depositing of data, and feedback as new features are implemented.

Enteropathogen Resource Integration Center

Bioinformatics Resource Center



ERIC will be a **pathogen-centric**, portal based system. ASAP (A Systematic Annotation Package for community annotation) from UW will be used to allow the scientific community to annotate genes for the five enteropathogens, and is online **now**.



Enteropathogen Resource Integration Center

Bioinformatics Resource Center

Welcome to ERIC BRC

Welcome to ERIC! - the Enteropathogen Resource Integration Center

ERIC is a new Bioinformatics Resource Center (BRC) for Biodefense and Emerging/Re-Emerging Infectious Diseases. Funded by the **National Institute of Allergy and Infectious Diseases (NIAID)**, ERIC will serve as a pathogen-centric resource for five members of the family *Enterobacteriaceae*:

- Diarrheagenic *E. coli*
- *Shigella*
- *Salmonella*
- *Yersiana enterocolitica*
- *Yersinia pestis*

Over the five year term of this contract, SRA International, teaming with the Genome Center of Wisconsin, will develop a Web portal system, which will integrate information on the genomic sequence, genome annotations, and related biological data for these five organisms. A key premise of ERIC is that it will be **pathogen-centric** – not centered around any one biological data type, but instead focused on the overall biology of these five enterobacteria.

Annotation information on these pathogens is already available through ERIC via GCW's community annotation application, ASAP. Over the next few years, this information will be expanded to include information on microarrays, proteomics, phenotypes, and genomic structures. ERIC will also utilize advanced text mining software and visualization tools to assist researchers in finding and interpreting the considerable amount of information that will be available.

Login to ERIC BRC

Username: admin

Password: *********

ASAP

ASAP is a tool for Community Annotation of Enterobacterial Genomes

ERIC's first component is the ongoing support and deployment of the existing sASAP (A Systematic Annotation Package for community analysis of genomes) application, for annotation of genomes for all five enterobacteria. ASAP is developed and maintained by Genome Center of Wisconsin (GCW). ASAP is built on open standards and is capable of accepting other pathogen data sets as the need or desire arises.

ASAP is operational now and is available for immediate use for querying, downloading and contributing to the annotation of the above enteropathogen genomes.

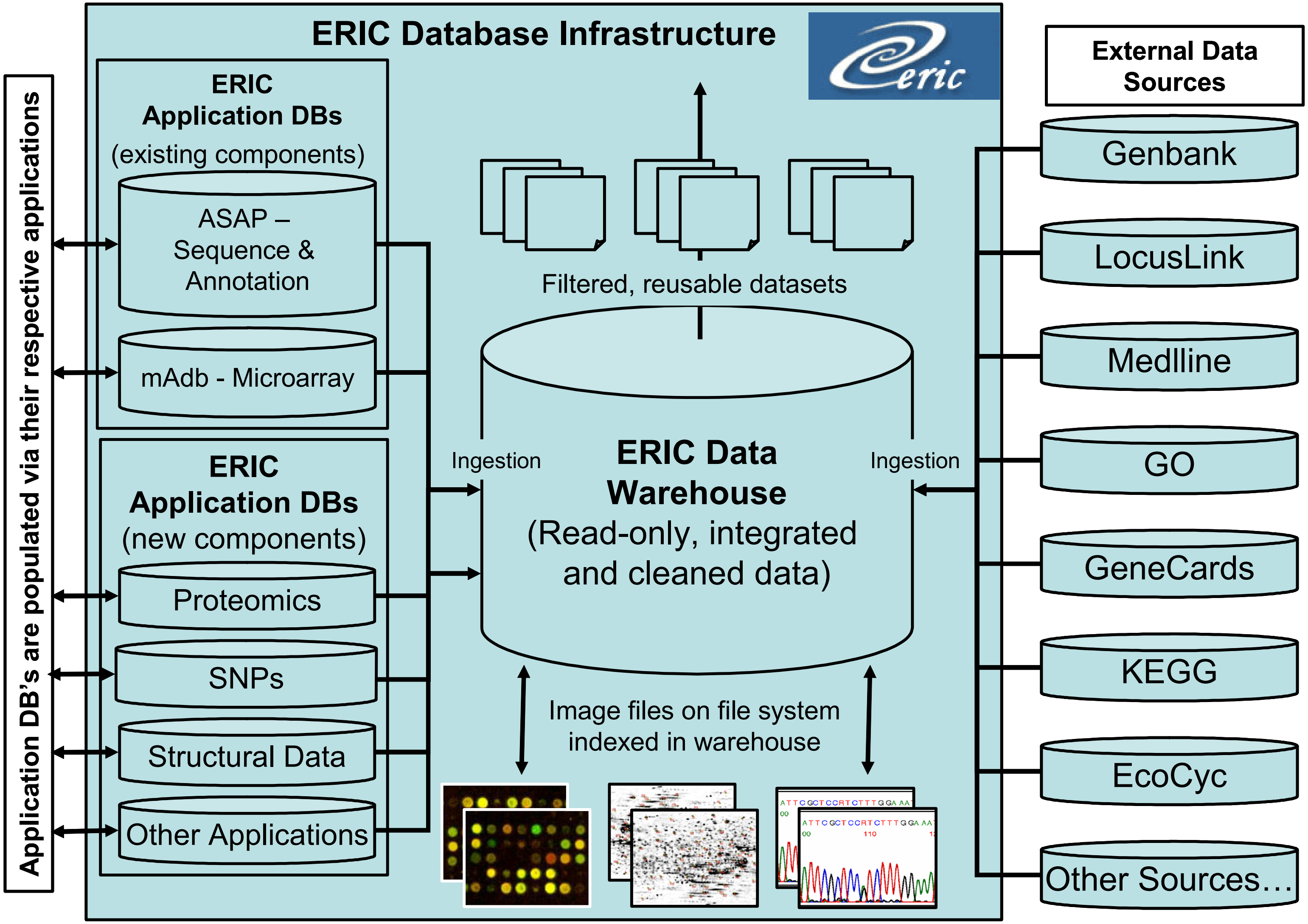
Further Information

- **Why Bioinformatics Resource Centers?**
- **The SRA/GCW Team and Our Approach**
- **NIAID Biodefense Web site**
- **NIAID BRC Program Web site**

For more information, please contact **info@ERICBRC.org**

This project has been wholly funded with Federal funds from the National Insitute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400040C

An important advantage of the ERIC portal design is that users will be able to **customize** it for their particular needs – and change it as their needs change.



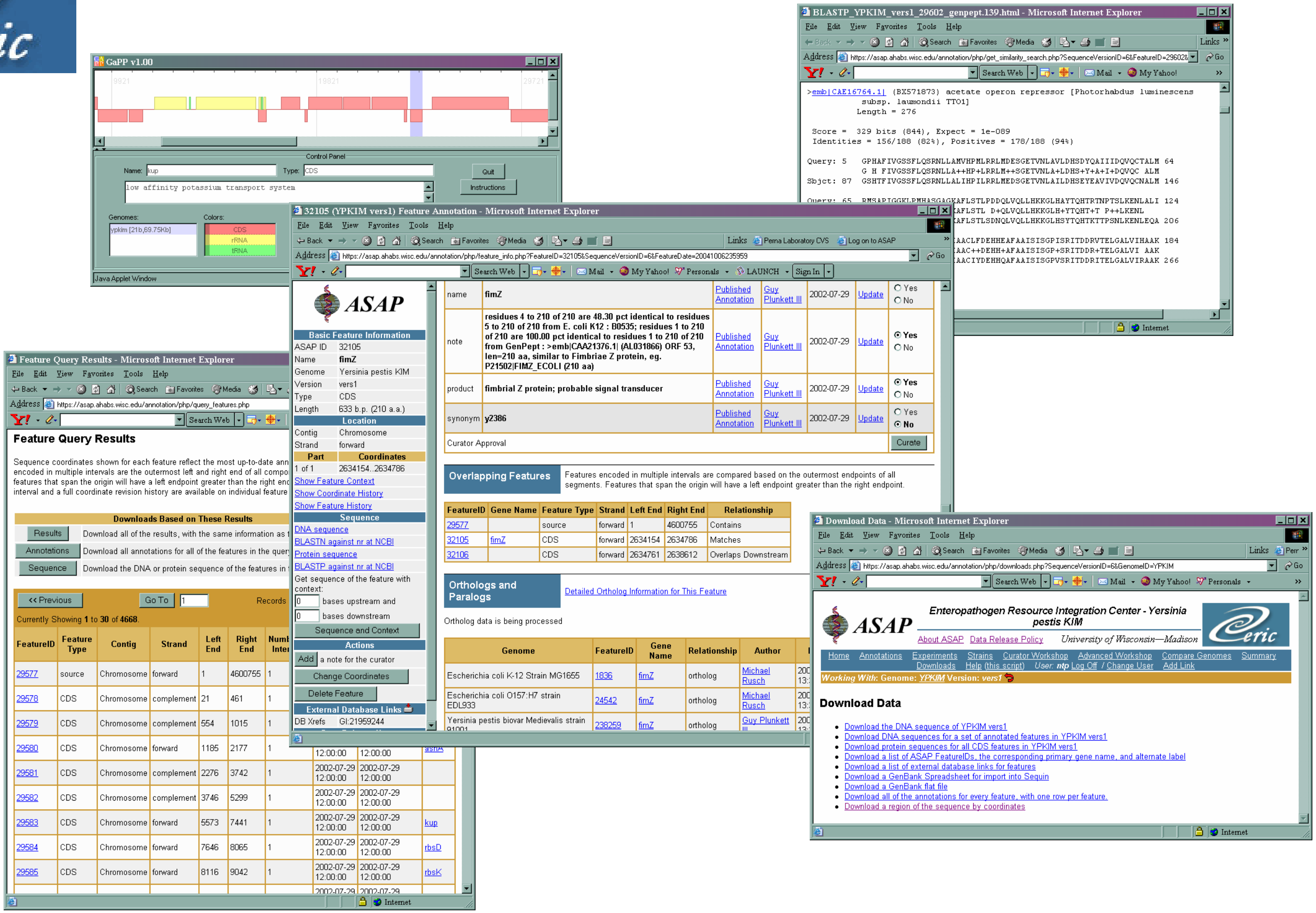
ERIC will make use of a **data warehouse** approach to store both contributed pathogen data (annotations, sequence, microarray, proteomics, etc.) and data from external sources. This will enable better integration of inputs from many sources.

System Hardware

- Sun - running Solaris OS
  - – SunFire V210s for computation
  - – 4 node computational cluster
  - – SunFire V480 for database server
- Oracle 10g Enterprise
- SAN – 3.5 TB expandable disk space
- L100 Tape Drive – 20 TB capacity

Development Methodologies

- Agile development
  - Daily Scrum
  - Short-cycle Iterations, once hardware installed
  - Close interactions with the enterobacterial research community, soliciting feedback on system development and features
  - Subversion for Configuration Management
  - Bugzilla for bug and change tracking
  - Peer review of software; formal QA and testing
- Focus on delivering scientist-friendly user interfaces that work together



Name: kup Type: CDS
low affinity potassium transport system
Genomes: ypkim [21b,69.75Kb]
Colors: CDS rRNA tRNA

BLASTP_YPKIM_vers1_29602_genpept.139.html - Microsoft Internet Explorer

Address https://asap.ahabs.wisc.edu/annotation/php/get_similarity_search.php?SequenceVersionID=6&FeatureID=29602&

```
>emb|CAE16764.1| (BX571873) acetate operon repressor [Photorhabdus luminescens
          subsp. laumondii TTO1]
         Length = 276

 Score =  329 bits (844), Expect = 1e-089
 Identities = 156/188 (82%), Positives = 178/188 (94%)

Query: 5   GPHAFIVGSSFLQSRNLLAMVHPMLRRLMDESGETVNLAVLDHSDYQAIIIDQVQCTALM 64
           G H FIVGSSFLQSRNLLA++HP+LRRLM++SGETVNLA+LDHS+Y+A+I+DQVQC ALM
Sbjct: 87  GSHTFIVGSSFLQSRNLLALIHPILRRLMEDSGETVNLAILDHSEYEAVIVDQVQCNALM 146

Query: 65  RMSAPIGGKLPMHASGAGKAFLSTLPDDQLVQLLHKKGLHAYTQHTRTNPTSLKENLALI 124
                             KAFLSTL D+QLVQLLHKKGLH+YTQHT+T P++LKENL
                             KAFLSTLSDNQLVQLLHKKGLHSYTQHTKTTPSNLKENLEQA 206

                             IAACLFDEHHEAFAAISISGPISRITDDRVTELGALVIHAAK 184
                             IAAC++DEHH+AFAAISISGP+SRITDDR+TELGALVI AAK
                             IAACIYDEHHQAFAAISISGPVSRITDDRITELGALVIRAAK 266
```

32105 (YPKIM vers1) Feature Annotation - Microsoft Internet Explorer

Links Perna Laboratory CVS Log on to ASAP

Address https://asap.ahabs.wisc.edu/annotation/php/feature_info.php?FeatureID=32105&SequenceVersionID=6&FeatureDate=20041006235959

ASAP

Basic Feature Information	
ASAP ID	32105
Name	**fimZ**
Genome	Yersinia pestis KIM
Version	vers1
Type	CDS
Length	633 b.p. (210 a.a.)

Location	
Contig	Chromosome
Strand	forward
Part	**Coordinates**
1 of 1	2634154..2634786

Show Feature Context
Show Coordinate History
Show Feature History

Sequence

DNA sequence
BLASTN against nr at NCBI
Protein sequence
BLASTP against nr at NCBI

Get sequence of the feature with context:
0 bases upstream and
0 bases downstream
Sequence and Context

Actions

Add a note for the curator
Change Coordinates
Delete Feature

External Database Links

DB Xrefs	GI:21959244

name	**fimZ**	Published Annotation	Guy Plunkett III	2002-07-29	Update	○ Yes ○ No
note	**residues 4 to 210 of 210 are 48.30 pct identical to residues 5 to 210 of 210 from E. coli K12 : B0535; residues 1 to 210 of 210 are 100.00 pct identical to residues 1 to 210 of 210 from GenPept : >emb|CAA21376.1| (AL031866) ORF 53, len=210 aa, similar to Fimbriae Z protein, eg. P21502|FIMZ_ECOLI (210 aa)**	Published Annotation	Guy Plunkett III	2002-07-29	Update	● **Yes** ○ No
product	**fimbrial Z protein; probable signal transducer**	Published Annotation	Guy Plunkett III	2002-07-29	Update	● **Yes** ○ No
synonym	**y2386**	Published Annotation	Guy Plunkett III	2002-07-29	Update	○ Yes ● **No**
Curator Approval						Curate

Overlapping Features

Features encoded in multiple intervals are compared based on the outermost endpoints of all segments. Features that span the origin will have a left endpoint greater than the right endpoint.

FeatureID	Gene Name	Feature Type	Strand	Left End	Right End	Relationship
29577		source	forward	1	4600755	Contains
32105	fimZ	CDS	forward	2634154	2634786	Matches
32106		CDS	forward	2634761	2638612	Overlaps Downstream

Orthologs and Paralogs

Detailed Ortholog Information for This Feature

Ortholog data is being processed

Genome	FeatureID	Gene Name	Relationship	Author	
Escherichia coli K-12 Strain MG1655	1836	fimZ	ortholog	Michael Rusch	200 13:
Escherichia coli O157:H7 strain EDL933	24542	fimZ	ortholog	Michael Rusch	200 13:
Yersinia pestis biovar Medievalis strain 91001	238259	fimZ	ortholog	Guy Plunkett III	200 13:

Feature Query Results - Microsoft Internet Explorer

Address https://asap.ahabs.wisc.edu/annotation/php/query_features.php

Feature Query Results

Sequence coordinates shown for each feature reflect the most up-to-date ann encoded in multiple intervals are the outermost left and right end of all compo features that span the origin will have a left endpoint greater than the right enc interval and a full coordinate revision history are available on individual feature

Downloads Based on These Results

Results	Download all of the results, with the same information as t
Annotations	Download all annotations for all of the features in the quer
Sequence	Download the DNA or protein sequence of the features in t

<< Previous Go To 1 Records

Currently Showing **1** to **30** of **4668**.

FeatureID	Feature Type	Contig	Strand	Left End	Right End	Numb Inter			
29577	source	Chromosome	forward	1	4600755	1			
29578	CDS	Chromosome	complement	21	461	1			
29579	CDS	Chromosome	complement	554	1015	1			
29580	CDS	Chromosome	forward	1185	2177	1	12:00:00	12:00:00	asnA
29581	CDS	Chromosome	complement	2276	3742	1	2002-07-29 12:00:00	2002-07-29 12:00:00	
29582	CDS	Chromosome	complement	3746	5299	1	2002-07-29 12:00:00	2002-07-29 12:00:00	
29583	CDS	Chromosome	forward	5573	7441	1	2002-07-29 12:00:00	2002-07-29 12:00:00	kup
29584	CDS	Chromosome	forward	7646	8065	1	2002-07-29 12:00:00	2002-07-29 12:00:00	rbsD
29585	CDS	Chromosome	forward	8116	9042	1	2002-07-29 12:00:00	2002-07-29 12:00:00	rbsK
							2002-07-29	2002-07-29	

Download Data - Microsoft Internet Explorer

Address https://asap.ahabs.wisc.edu/annotation/php/downloads.php?SequenceVersionID=6&GenomeID=YPKIM

ASAP *Enteropathogen Resource Integration Center - Yersinia pestis KIM*

About ASAP Data Release Policy *University of Wisconsin—Madison*

Home Annotations Experiments Strains Curator Workshop Advanced Workshop Compare Genomes Summary Downloads Help (this script) *User:* ***ntp*** Log Off / Change User Add Link

Working With*: Genome: *YPKIM* Version: *vers1

Download Data

- Download the DNA sequence of YPKIM vers1
- Download DNA sequences for a set of annotated features in YPKIM vers1
- Download protein sequences for all CDS features in YPKIM vers1
- Download a list of ASAP FeatureIDs, the corresponding primary gene name, and alternate label
- Download a list of external database links for features
- Download a GenBank Spreadsheet for import into Sequin
- Download a GenBank flat file
- Download all of the annotations for every feature, with one row per feature.
- Download a region of the sequence by coordinates

ASAP meets needs for direct, community-wide input (with authorship and annotation history tracking), multiple annotations of features, evidence codes, using controlled vocabularies, curatorial review, support of cross-genome comparisons, and web-based updating and access. In addition, ASAP links annotations to features, and not specifically to genome assemblies.

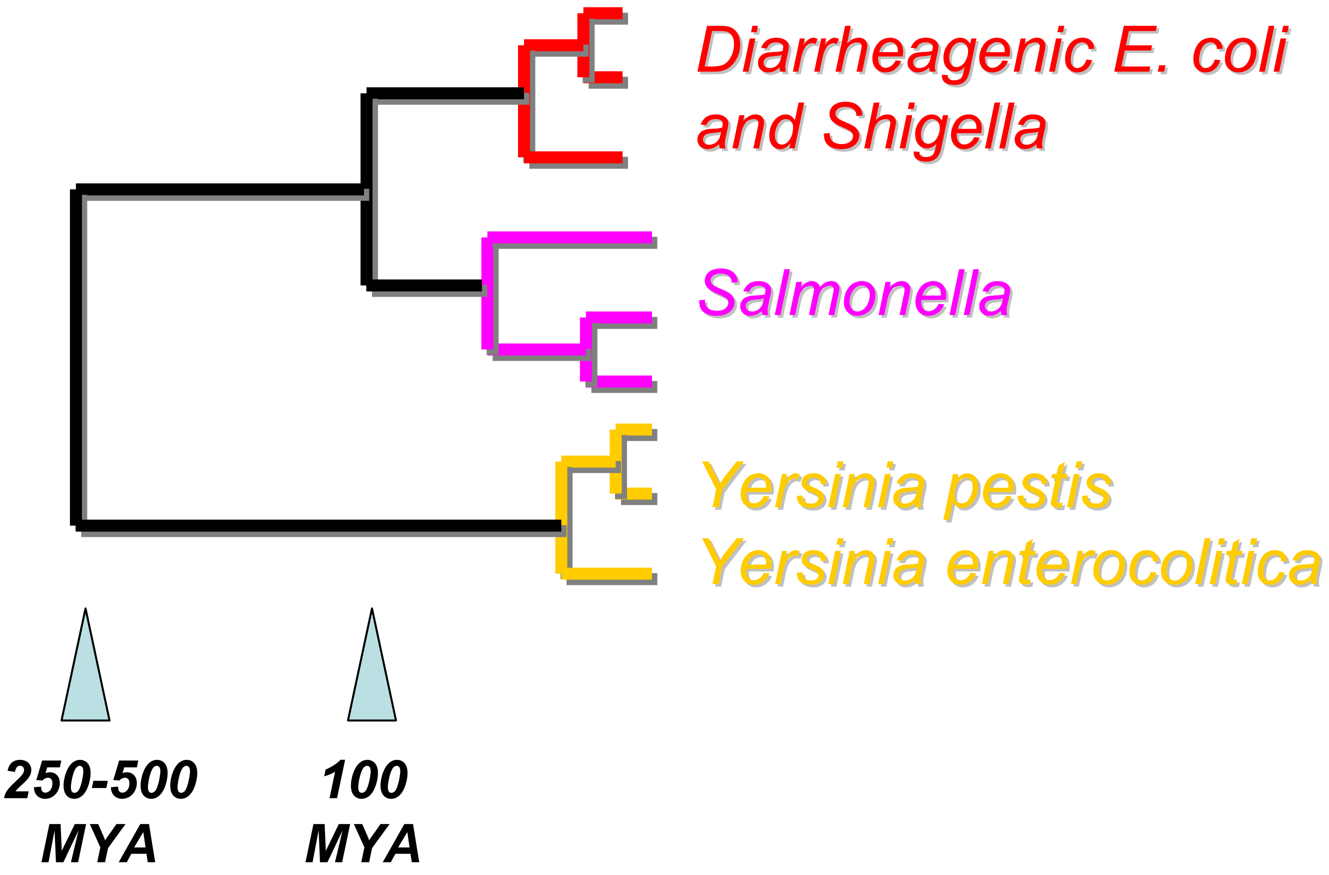
ERIC Genomes Available or Expected

Organism	Complete*	Ongoing	Expected
E. coli and Shigella	4	4	10
Salmonella	3 (3)	10	10
Yersinia	3	1	10
subtotal	10 (3)	15	30

*** Draft 4x coverage**

Total over next two years: ~ 60 genomes

ERIC Enterobacteria



Diarrheagenic E. coli and Shigella

Salmonella

Yersinia pestis

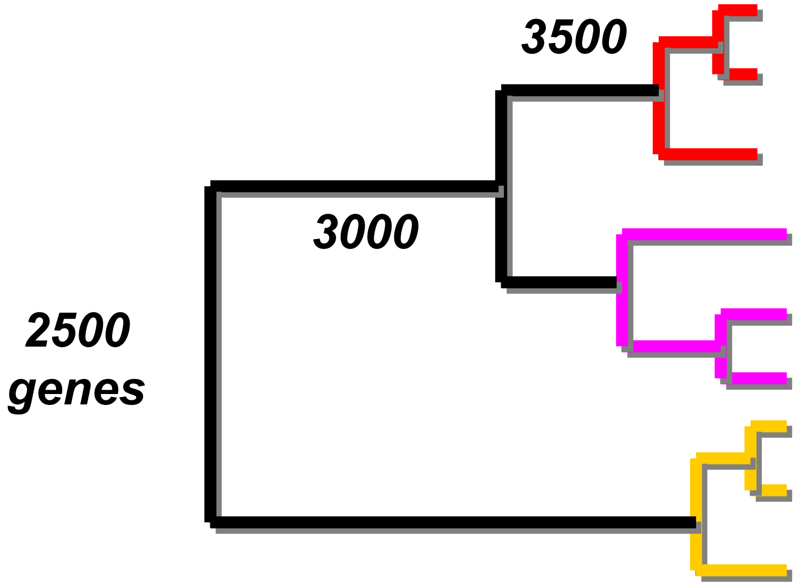
Yersinia enterocolitica

250-500 MYA

100 MYA

ERIC Enterobacteria

Unique to each strain



3500

3000

2500
genes

200-1500

Diarrheagenic E. coli and Shigella

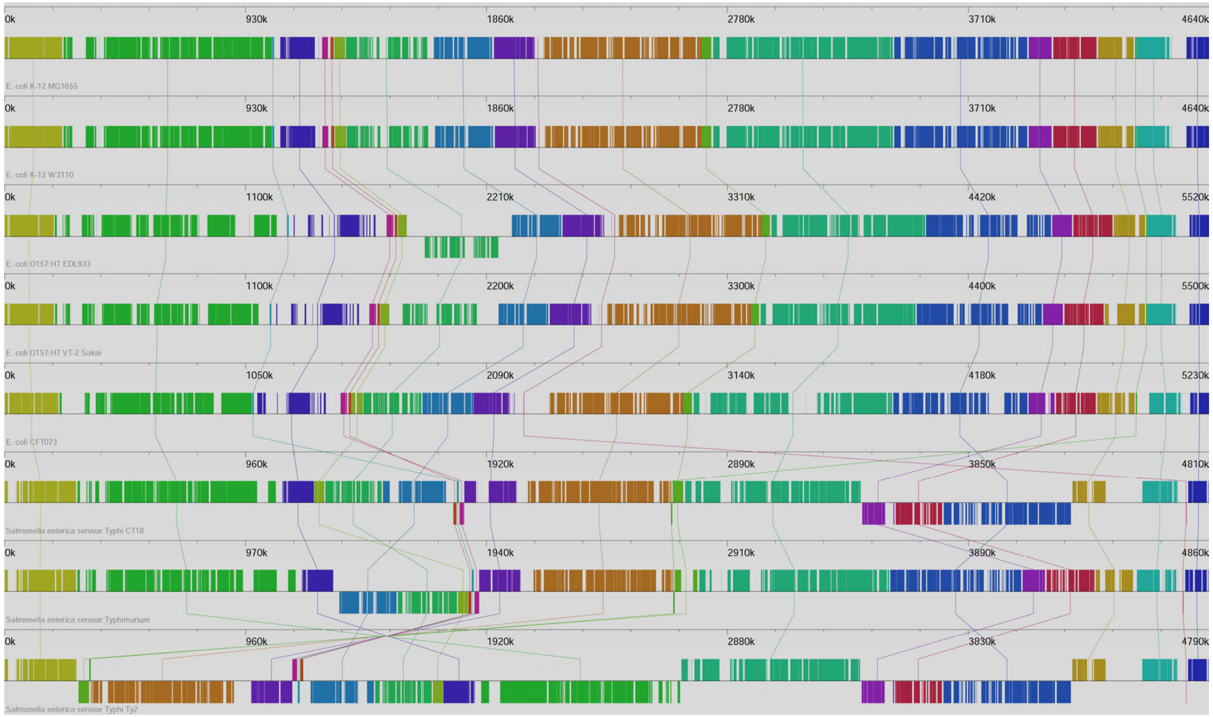
500-1000

Salmonella

<100
?

Yersinia pestis
Yersinia enterocolitica

Total Number of Gene Products ~32,000



7111778-028

- ERIC will contain a number of tools for **comparative genomics**, such as Mauve, which has the distinct advantage of allowing alignment of more than two genomes, as well as being able to handle chromosomal rearrangements

- Also, **functional analysis** tools – MatchMiner (batch-translates among many types of gene and protein identifiers) and GoMiner (leverages the Gene Ontology (GO) to identify the biological processes, functions and components represented in a list of genes)

Annotation/Curation Strategy

- Leverage relationships among sequences
  - Propagate annotations across orthologs
- Balance automated and manual efforts
  - Curatorial review of both for quality control
- Task-based approach
  - Topical rather than linear
  - Examples: O-antigen, Invasion, Toxins

Annotation/Curation Strategy

- High Priority Tasks
  - – Establishing Orthology
  - – Polymorphisms useful for diagnostics and subtyping
  - – Gene products associated with pathogenesis
  - – Fixing errors

ERIC Annotation Types

- Gene name
- Identifiers
- Product names
- Function
- EC Numbers
- Notes

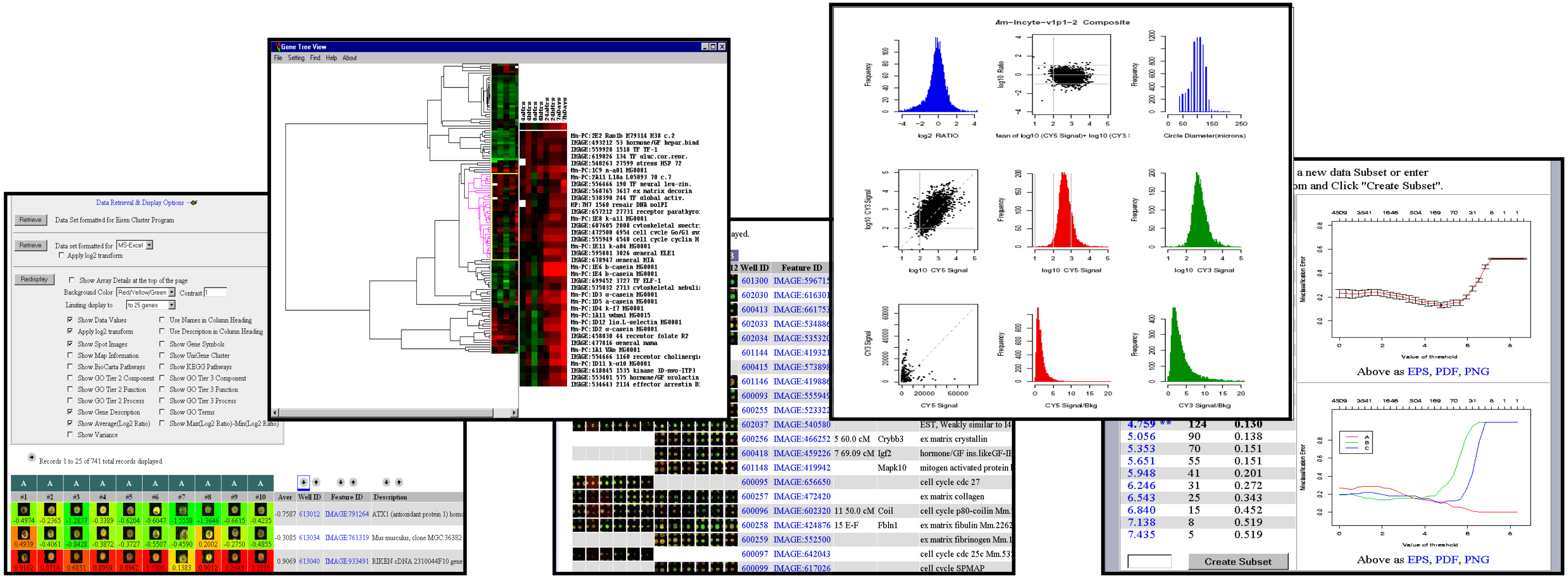
- GO
- Mutant phenotypes
- Over expression phenotypes
- Genetic Interactions
- Molecular Interactions
- Biochemical properties
- Protein modifications
- Regulation
- Structure
- Comments about literature

Ontologies

- Features
  - Expanded INSD (GenBank, EMBL, DDJB)
  - Examples (CDS, promotor, rRNA, tRNA, …)
- Feature Qualifiers
  - Expanded INSD (GenBank, EMBL, DDJB)
  - Examples (/gene, /product, /note, /function,…)
- Gene Function
  - MultiFun
  - GO
- Evidence

Enteropathogen Resource Integration Center

Bioinformatics Resource Center



For **microarray** work, ERIC will incorporate the mAdb system, built for the intramural program of the National Cancer Institute. This web-based system is supporting over 1,200 users and their collaborators worldwide and over 40,000 arrays. It incorporates numerous tools for filtering data, providing statistical analysis, and allowing users to visualize the data.

SRA Text Mining Technology

- **SRA is an industry leader in natural language processing (NLP)-based text mining**
  - Dedicated group of linguists and software engineers
  - Routinely win Government text mining competitions (e.g. Message Understanding Competitions (MUC))
- **Extensive experience in multilingual information extraction, text clustering, and text summarization – this is not keyword searching**
- **Numerous commercial and government clients/applications**
  - Health care organizations (fraud detection); Financial services (anti-money laundering, e-mail surveillance); Government (homeland security, e-Government, business intelligence)

SRA Text Mining - GeneTag

- **GeneTag prototype successfully tested in bioinformatics to mine scientific literature for the CDC and for the American Cancer Society.**
- **The prototype incorporates the following functions:**
  - □ GeneTag automatically derives gene annotations from MEDLINE abstracts, including the functions of a gene, as well as the diseases, tissues, and other genes associated with it.
  - □ The system output for each gene is a structured summary of the above, hyperlinked to the abstracts from which they are derived.
  - □ GeneTag addresses the serious problem of varying gene nomenclature by automatically linking the various ways of referring to a specific gene found in the literature.
- **Text mining tools can be used for many of ERIC’s goals** – function, biological networks and pathways, and certainly the identification of factors for virulence, infectivity and pathogenicity.
- **Discovering heretofore unrecognized relationships in the literature** may be key to designing experiments which will lead to the identification of targets for vaccines, therapeutics, and diagnostics.

ERIC Team:

Scientific Co-Directors:

John Greene, Ph.D. - PI and Project Director
Nicole Perna, Ph.D. - Scientific Co-Director
Fred Blattner, Ph.D. - Scientific Co-Director

ERIC Curators:

Guy Plunkett III, Ph.D. - Senior Curator; David Bowen, Ph.D.; Val Burland, Ph.D.; Eric Cabot, Ph.D.; Jeremy Glasner, Ph.D.

ERIC Technical Team:

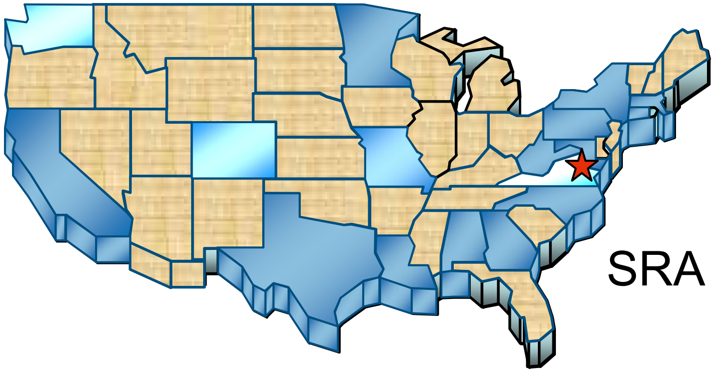
Matt Shaker; Robin Martell; Tom Hampton; Lorie Shaull; Panna Shetty
Paul Liss; Michael Rusch

Scientific Co-Directors

- **John Greene, Ph.D.** – PI and Project Manager
  - Ph.D., Harvard, Genetics; B.S., MIT ’83; IS Certificate from GWU
  - Eight years bioinformatics experience - HGS, Gene Logic, and SRA
  - Discovered 9 novel genes at HGS; built microbial database system for Pharmacia/Upjohn collaboration
- **Nicole Perna, Ph.D.** – Asst. Professor, UW
  - Sequenced numerous enterobacterial genomes
  - Comparative Genomics of Enterobacterial Pathogens & Evolution of Virulence Determinants
  - Developed ASAP bioinformatics system for community annotation
- **Fred Blattner, Ph.D.** – Smithies Professor of Genetics
  - Founding Director of the Genome Center of Wisconsin
  - Founder and President, DNASTAR, Inc.
  - Co-founder of Nimblegen Systems, Inc.

SRA Profile & Bioinformatics

- Information Technology/ Systems Integration company based in Fairfax, Virginia
- Founded in 1978; 26 consecutive years of growth and profitability
- FY2004 revenues of $616M
- Public company on NYSE - SRX
- Over 3,500 IT professionals
- SRA essential attributes
  - Ethic of honesty and service
  - Quality work and customer satisfaction
  - People orientation



SRA Offices

Biomedical Informatics Projects

Microarray Database (mAdb) – **NCI, NIAID, NIMH, NHGRI, FDA-CBER, GIS, NKI**

- End-to-end support for microarray data analysis (mAdb)
- Analysis tool & database development and support

NCI

- Open Source functional genomics tools (GoMiner, MatchMiner)
- Support for microarray statistics

NINDS

- Clinical Trial and Protocol Data Management support
- Bioinformatics Core staff

CIT

- Microarray algorithm development, data analysis, & visualization in a high-performance and parallelized computing environment
- Archiving of Medical Images
- Biostatistics support services
- Bioinformatics development, operational, and maintenance services

eric

http://www.ericbrc.org

info@ericbrc.org